图书情报工作
第 64 卷 第 7 期 2020 年 4 月
ChinaXiv合作期刊

文化遗产语义组织研究进展*

■ 李章超 何琳
南京农业大学信息管理系 南京 210095

摘 要：[目的/意义]针对文化遗产语义组织发展现状展开研究，对我国文化遗产研究具有重要参考价值。[方法/过程]采用系统调研法、案例分析法和统计分析法，以调研数据概括为基础，从语义组织方式和知识服务与工具两个方面对文化遗产项目语义组织研究现状进行梳理，从知识建模、知识抽取和知识挖掘与利用三个维度对文化遗产语义组织关键技术进行剖析。[结果/结论]研究发现，数据互操作、领域本体标准化、个性化语义、自动化工具和数据版权是未来文化遗产语义组织发展的关键。

关键词：文化遗产 语义组织 本体技术 知识服务
分类号：G255
DOI：10.13266/j.issn.0252－3116.2020.07.001

chinaXiv:202304.00292v1

1 引言

文化遗产是人类在社会历史发展过程中创造的具有历史、艺术、科学等价值的文化财富，深入研究和挖掘文化遗产，有利于阐发文化精髓，保护与传承文化遗产，推动中外文化交流互鉴。随着信息化技术的发展，越来越多的文化遗产实现了数字化，文化遗产资源规模日益庞大，种类繁多，为文化遗产的保护、研究和传承积累了良好的资源基础。然而由于这些资源的数据存储形式和格式多样，缺乏统一的描述标准，使得文化遗产资源数据表现出较强的异构性。如何将大量异构的文化遗产资源以计算机可处理及可理解的方式组织起来，成为当前亟待深入研究和解决的问题。

近年来，数字人文及人文计算研究得到了计算机、图书情报和历史学领域学者的广泛关注，对于文化遗产领域数字资源的信息组织，学者们尝试将智能信息处理技术与文化遗产资源研究结合起来，开展文化遗产语义组织的理论和实践研究。借助语义网及机器学习等技术，学者们对图书馆、博物馆及档案馆的典藏资源进行了知识抽取、知识组织及知识服务研究，逐步积累了文化遗产信息资源管理的相关理论、方法和技术。在此背景下，本文以国内外典型文化遗产项目为基础，对国内外文化遗产语义组织研究现状展开研究，以期为文化遗产资源的语义组织提供借鉴。

2 文化遗产项目语义组织研究现状

随着人文大数据的发展，文化遗产数据经历了数字化、结构化、语义化的发展，逐步探索文化遗产中蕴含知识点间更深层次的语义关系。本文以调研数据概括为基础，从语义组织方式和语义知识服务与工具两个方面，对在技术、规范和系统化程度等方面具有代表性的文化遗产项目的语义组织现状进行研究。基于各文化遗产项目平台的资料，本文对典型文化遗产项目的研究主题、来源资源类型、语义组织模式和提供的语义服务进行总结，见表 1。

2.1 调研数据来源概括

（1）文化遗产项目主要研究重点。国内外典型文化遗产项目主要围绕历史人物、历史事件和文化收藏等主题开展语义组织研究。一些学者针对历史人物设计语义描述模型，如 CBDB（中国历代人物关系数据库）围绕中国历史人物传记资料展开研究，涉及亲属关系、社会关系等实体[1]；宋代学术语义网络平台对宋代人物之间的学术传承关系和亲属关系进行组织和重构；历代进士登科数据库针对公元 6 世纪以来 10 万余历代进士登科人物的登科资料展开研究。还有一些语义门户围绕历史事件展开研究，如芬兰文化遗产项目

* 本文系国家社会科学基金项目“基于典籍的中华传统文化知识表达体系自动构建方法”（项目编号：18BTQ063）研究成果之一。
作者简介：李章超（ORCID：0000－0002－9252－2142），博士研究生；何琳（ORCID：0000－0002－4207－3588），教授，博士，博士生导师，通讯作者，E-mail：helin@njau.edu.cn。
收稿日期：2019－07－10 修回日期：2019－11－10 本文起止页码：4－12 本文责任编辑：杜杏叶

表 1　典型文化遗产项目概况

项目名称	研究主题	来源资源类型	语义组织模式	语义服务
中国历代人物关系数据库(CBDB)	中国历史人物传记资料	正史列传、墓志铭、墓表、地方志列传、人物传记索引	数据库;以人物关系为中心	结构化查询(入仕、社会关系、职官、亲属关系、关系网络、配对关系)、群体传记学分析、社会网络分析、地理系统分析
Europeana	欧洲图书馆、博物馆、档案馆馆藏	书籍、音乐、艺术品等数字遗产资源	知识图谱;以文化收藏为中心	精准化搜索与过滤工具、开放 REST API 接口
WarSampo	第二次世界大战(芬兰)	战争资料、战争日记、战争相册集、战争回忆录等资料	关联数据;以战争事件为中心	数据库全局下载、URI 重定向、关联数据浏览与查询、数据生产、编辑与认证、可视化分析
MuseumFinland	芬兰博物馆馆藏	芬兰各博物馆、诺基亚和 TietoEnator 等公司等机构的收藏	关联数据;以机构收藏为中心	基于内容的智能搜索和浏览服务
宋代学术语义网络平台	宋代人物间学术传承关系和亲属关系	CBDB 中的宋代人物数据	知识图谱;以人物关系为中心	知识图谱结构探索、RelFinder 关系发现
历代进士登科数据库	登科人物的传记资料	传世文献、出土史料	数据库;以人物为中心	搜索浏览、数据统计、可视化分析

WarSampo 对战争日记、战争相册集、战争回忆录等战争资料中与芬兰有关的第二次世界大战事件展开研究[2]。还有部分语义门户以文化收藏为核心进行数据关联,主要包括藏品、绘画、音乐、建筑等,如芬兰博物馆项目 MuseumFinland 围绕芬兰各博物馆和诺基亚等公司的收藏展开研究[3],欧洲文化遗产项目 Europeana 对欧洲多个图书馆、博物馆、档案馆等机构的书籍、音乐、艺术品等馆藏资源展开研究[4]。

(2)文化遗产项目数据来源。文化遗产资源是文化遗产项目的基础和关键。通过对各文化遗产项目网站平台的数据和相关资料进行调研,发现当前国内外文化遗产项目多以公共文化机构、第三方机构和高校、科研院所等机构的馆藏或数字资源为基础,主要可分为自建关系数据库和多源异构数据库。

自建关系数据库,主要是指文化遗产项目通过自身搜集、翻阅、辑录原始史料或者利用二手史料构建数据库。自建关系数据库能够保证数据的原创性和真实性,但是需要进行大量的人工处理,效率较低。如 CBDB 项目早期对正史列传、墓志铭、墓表、地方志列传及人物传记索引等原始语料进行手工处理[5],历代进士登科数据库中的数据多是由龚延明教授及其团队手工辑录的。

多源异构数据库,主要实现高校、科研院所、公共文化机构、出版社等其他组织构建数据库的集成。多源异构数据库的数据获取方便,但是呈现较强的多源异构性,重点在于实现不同类型、组织方式的数据或数据库之间的兼容。如 WarSampo 项目中的文化遗产数据多来源于芬兰各博物馆和档案馆[2],CBDB 中部分数据来源于出版社发行的人物传记索引、年表以及麦吉尔大学的明清妇女著作数据库、中央研究院的明清档案人名权威数据库等[5]。

2.2　语义组织方式

语义组织方式是根据文化遗产资源的特点,以某种方式实现文化遗产资源的有序化、规律化或者系统化,主要经历了非结构化、结构化、关联化和智慧化四个阶段,实现文化遗产数据资源从文献数据到量化数据、智能数据的发展。

(1)非结构化阶段。文化遗产数据最初基本都以非结构化方式存在于各类信息资源中,表现出极强的不规则性和不完整性。在文化遗产数据组织的早期阶段,主要通过包括 OCR 识别和人工录入的方式实现文化遗产资源的数字化。这一阶段需要耗费大量的时间、精力阅读和整理文化遗产资源,同时需要历史研究人员对资源进行主观的归纳与演绎。比如,历代进士登科数据库使用手工方式进行文献资料的搜集和录入,构建出相应的《中国历代登科总录》《宋代登科总录》《明代登科总录》等索引资料[6]。

(2)结构化阶段。文化遗产组织的结构化阶段是指利用自然语言处理、元数据等技术方法,对文化遗产资源进行典籍分词、词性标注、命名实体识别等技术处理,实现文化遗产资源的资料化、规模化、有序化和规律化。比如,CBDB 利用正则表达式等计算机文本挖掘技术,从数字文化资源中精准地抽取出历史人物资料[7]。这一阶段实现了分词及词汇级的文本抽取工作,产生了了大量的主题索引,如《十三经索引》《二十四史地名索引》等,以及大量专科词典,如《中国历史地名大词典》《历代职官词典》等研究成果。

(3)关联化阶段。文化遗产组织的关联化阶段主要是指在结构化数据的基础上,构建本体框架模型,实现概念关系抽取和语义关联。关联化阶段是目前大部分文化遗产项目建设的目标,依据文化遗产不同的研究领域、研究问题、研究情境以及史料特点,构建主题

chinaXiv:202301.00202v1

ChinaXiv合作期刊

词表、领域本体等概念关系描述工具，在此基础上形成了关联数据，为实现文化遗产资源时间、空间、主题等分析提供了关联化的数据。如，WarSampo 构建二战本体框架模型和关联开放数据云，实现接入云的组织和个人获取、共享结构化文化遗产资源数据，形成文化遗产数据之间的相互语义关联[2]；MuseumFinland 构建符合博物馆馆藏特征的本体模型，实现芬兰各博物馆的馆藏数据的语义关联[3]。

（4）智慧化阶段。文化遗产组织的智慧化阶段是指利用前述阶段构建的数据，从时序、空间、主题、网络等多个维度进行更深层次的语义分析。智慧化阶段逐渐形成了数据驱动的数字人文研究新范式，即在关联化数据的基础上，利用新工具和新方法对旧问题进行新思考，同时根据现有的大量数据提出新问题。关联化阶段是目前部分文化遗产项目努力达成的目标，宋代学术语义网络平台基于 CBDB 中的数据，对宋代人物进行时间序列分析、空间地理分析、社会网络分析等可视化分析。陈佩诗基于《明清台湾行政档案》与《古契约文书》的数据，通过对清代行政文书在不同行政部门流转的引用情况，进而分析清代行政处理效率及流程[8]。

2.3　语义知识服务与工具

语义知识服务是以文化遗产项目中的数据为基础，按照用户的需求特点，利用系统提供的语义知识服务与工具，如精准化、层级化的知识检索和浏览服务，自动化、关键词式或人工化的标记与分词工具，基于时间、地点、人物等特征的字频、词频统计分析工具，以及基于时间、GIS 地理数据分析及空间分析、人物关系等维度的可视化分析工具或服务，从而有针对性地提取相应的知识，搭建知识网络，为用户提供符合语义的知识内容或解决方案。调研发现，国内外典型的文化遗产项目均开通相对应的语义门户或系统，为用户提供层级化、可视化的信息浏览、检索服务；文化遗产项目还提供概念间关系的语义检索等关联数据服务，比如宋代学术语义网络平台设计了学生关系、学术传承关系等 39 种人物关系[9]，能够清晰地揭示出人物概念之间的多层关系，为数字人文研究提供数据支撑。同时，多数项目基于知识图谱和关联数据等技术，开发相应语义工具，为用户提供语义检索、语义关联、知识发现和语义信息可视化等个性化语义知识服务[10]。

语义知识服务方面，MuseumFinland 为用户提供分面搜索和语义链接、浏览服务。历代进士登科数据库在实现“检索浏览”和“分类导航”功能的基础上，为用户提供姓氏统计、朝代统计等多维度的统计功能[11]。WarSampo 为用户检索提供不同类型的战争信息透视图，同时为用户提供相关数据推荐，具体包括数据库全局下载、URI 重定向、关联数据浏览、SPARQL 查询以及数据生产、编辑、认证和信息可视化等多项个性化知识服务[2]。

语义工具方面，CBDB 开发 CBDBRegexMachine 工具，基于正则表达式为用户从数据库中挖掘大量知识提供便利，帮助研究者实现中国历史人物传记数据的挖掘和可视化。Europeana 提供精准化搜索与过滤工具，开放 REST API 接口，帮助用户快速找到所需内容，同时允许开发人员使用数据库中的数据进行应用开发。宋代学术语义网络平台构建“知识图谱结构探索”和“关系发现”工具，通过可视化的方式帮助用户了解宋代学术网络的结构和关系，同时支持用户自主探索和发现实体关系[9]。

3　文化遗产语义组织关键技术分析

语义网及机器学习技术等的发展，为文化遗产资源知识体系构建、知识融合和知识应用等问题提供了有效解决方案，能够将文化遗产从资源服务层面提升为计算机可处理可理解的知识服务层面。具体来说，涉及到知识建模、知识抽取和知识挖掘等方面的技术与方法。

3.1　知识建模

文化遗产资源规模庞大、结构混乱，表现出较强的异质性、异构性。知识建模，是一种结构化、模型化的知识表达方式，能够实现文化遗产知识的结构化、语义化和共享化，为知识服务提供重要支撑。传统的知识建模主要以分类法、叙词表等为主，随着语义网技术的发展，越来越多的文化遗产项目选择使用元数据及本体技术进行知识建模。

3.1.1　叙词表

20 世纪 60 年代，叙词表迅速发展，涵盖了各个领域。文化遗产领域，Getty 研究所在国际标准的基础上，系统地构建人文领域词表，涉及艺术、建筑、书目、档案等多个主题，包括构建艺术和建筑叙词表（AAT）、保护叙词表（CT）、文化对象名称规范表（CONA）、地理名称表（TGN）、艺术家联合表（ULAN）和图像规范表（IA）[12]。此外，还有部分语义门户针对特定项目或领域构建叙词表，如 CBDB 构建中国古代官名表和地址表，芬兰国家图书馆构建涉及文化遗产、艺术、健康等各个研究领域的芬兰通用叙词表（YSA）[13]，美国国会

chinaXiv:202304.00092v1

图书馆针对馆藏构建美国国会图书馆主题词表(LCSH),武汉大学数字人文研究中心构建敦煌壁画主题词表(DMT)[14]等。

3.1.2 本体建模

本体模型主要包括事件本体、人物本体和人物－事件本体3种类型。其中,事件本体以事件类的格结构作为主线,主要包含对象要素、动作要素、时间要素、环境要素、断言要素和语言表现要素[15],如芬兰 MuseumFinland 基于大事年表创建历史事件本体,实现收藏品、材料、人物、位置、时间和收藏机构等元数据在相关事件的互操作,对实体的不同状态进行描述,实现文化遗产之间的语义关联[3]。人物本体以人物为知识组织主线,揭示人物之间的社会关系,实现人物关系的形式化和结构化[16],如 CBDB 围绕中国古代人物构建本体模型,BiographySampo 构建分类人物本体模型。人物－事件本体将人物和事件进行关联用作主线,如 WarSampo 将事件类加入人物本体中,将在形式上和数量上有较大差距的人物信息协调成为一系列事件,增强了模型的扩展性。

文化遗产领域常用 CIDOC CRM[17]、EDM(Europeana Data Model)[18]、BIBO[19]、HOPE[20]、SEM[21]等语义模型作为资源内容的描述规则,将不同形式的元数据映射到通用的底层本体模型上,在此基础上构建能够揭示数字遗产资源不同概念间关联关系的本体模型,并利用最小元数据模式对事件进行精准识别和理解,达到事件消歧的效果[13],从而进行事件知识表示。其中,国际文献工作委员会构建的概念参考模型 CIDOC CRM 是目前文化遗产领域规模最大、标准化程度最高的本体框架之一,能够有效促进文化遗产信息源的集成、转接和相互交换,目前在芬兰二战语义门户 WarSampo、Getty 和世界遗产基金会构建的文化遗产门户 Arches[22]、英国博物馆的 ResearchSpace[23]、古典艺术遗产门户 CLAROS[24]等文化遗产项目得到广泛使用。欧洲数字人文资源整合项目 Europeana 构建了 EDM 模型,对类和属性进行定义,揭示了聚合结构关系、资源对象间关系、事件情景关系和资源对象主题关联关系4种资源关联关系,在为其他机构提供基础模型框架的同时,也在不断扩充自身的数字内容。基于 EDM 模型,欧洲乐器博物馆联合创建的 MIMO 项目实现了乐器的集成[25],CARARE 实现考古和遗产领域数字内容的集成[26],伦敦国王学院的 SPQR 项目实现6.8万余件古希腊、古罗马碑文、铭文资源的集成。

世界范围内,本体框架模型数量多、各成规范,同时文化遗产的内容、来源、语言、格式和标准多样,无法避免地给数据的语义集成造成一定的困难;同时,在国内文化遗产领域,目前多是基于项目或朝代构建本体模型,尚无像 CIODC CRM、EDM 一类的标准化强、通用性高、系统化的本体框架模型,更没有针对中国传统文化构建的本体框架模型,这应该引起未来数字人文研究与实践的关注与重视。

3.1.3 元数据

目前,元数据是绝大多数文化遗产项目实现历史文本(非结构化)到数据(结构化)的通用解决方案,能够实现来源于不同数据集的异构数据的共享、交互和整合[27]。目前业界关注的重点在于元数据的语义互操作和元数据标准。

元数据的核心是解决异构数据的语义互操作问题,主要包括元数据扩展和元数据对齐两种思路。元数据扩展,即对现有元数据进行扩展,比如 WarSampo 基于 CIDOC CRM 框架模型实现元数据的管理与扩展;意大利文化遗产项目 Protocollo Informatico 基于 DC 元数据模型进行元数据元素拓展,同时实现项目模型与 DC 元数据元素的映射。元数据元素对齐[28],比如 MuseumFinland 将异构数据对齐、转化为符合 DC 标准的元数据,实现文化遗产资源的元数据表示。

元数据标准是元数据的基础和关键[29],使用用户熟知、易操作的元数据标准可以帮助实现数据的对齐与合并。当前文化遗产领域常使用元数据标准 DC(Dublin Core,DC)的命名空间及其向下拓展原则对数据库中收集对象的元数据的元素集合、术语和属性进行定义[30],在图书馆、政府网站等领域得到广泛使用。文化遗产领域,Europeana、WarSampo 和 MuseumFinland 等项目均以 DC 元数据标准为基础进行数据映射与对齐。另外,还有档案馆常用的 EAD 标准和博物馆的 LIDO 标准等来实现不同元数据模型元素之间的映射。构建文化遗产资源的框架模型,通常使用 W3C 推荐的 RDF(S)和 OWL 等语义网技术标准,并利用 SKOS 来构建资源描述所需的受控词汇表[31]。

3.2 知识抽取技术

在文化遗产领域,知识抽取主要是识别不同文化遗产资源中蕴含的知识点及其之间的语义关系。目前在知识抽取领域取得了较大的进展,学者们将这些技术和方法应用于文化遗产领域信息资源的知识抽取中。根据处理的信息资源对象的差异,规则匹配与机器学习技术是目前被广泛使用的方法。

chinaXiv:202304.00292v1

ChinaXiv合作期刊

3.2.1 规则匹配方法

规则匹配的方法以领域知识为基础，通过人工对文本特征进行分析，构建相对应的规则，并编写正则表达式，从而实现对文本的模式匹配，最终实现基于规则的知识抽取。这类方法主要适用于带抽取的语料在句法上具有一定的内在规律的文本，例如，典籍、地方志等。从实践上来看，CBDB 项目的历史学专家和计算机专家针对历史人物的特征设计相应的正则表达式，同时由编辑团队对匹配到的文本进行核验[7]，针对地方志、墓志铭本文的行文特点，设计了有针对性的知识点抽取规则，规则匹配法运用广泛，技术相对成熟，但是由于其规则的制定存在一定的主观性，无法保证知识的逻辑性、系统性和完整性；同时，由于规则匹配法的针对性较强，导致其在不同的文本、不同领域间的移植性较差。对于该问题，一些项目尝试利用计算机自动学习信息抽取规则。

3.2.2 机器学习的相关方法

在文化遗产领域的知识抽取工作中，目前常用的机器学习方法包含基于特征的方法和基于神经网络的方法。历史学专家可以通过约定的规则对训练文本进行标注，形成训练语料库，建立相应的训练语料模型。通过训练语料的学习，系统能够对新的文本进行处理。在历史学领域，以特征向量为基础的机器学习方法能够取得较好的效果，具有一定的代表性。

自然语言处理技术也常被用于文化遗产语义组织的研究中，能够实现文化遗产知识的文本分类、自动分词、命名实体识别、依存句法分析、事件抽取等功能，常用的算法模型包括支持向量机、决策树、随机森林等传统机器学习算法，条件随机场等注重前后特征的序列标注算法以及目前运用广泛的卷积神经网络、循环神经网络的深度学习技术。比如，意大利自然语言处理实验室基于文化遗产项目 CHROME 的开发需要，创建了 LinguA（Linguistic Annovation pipeline）、READ-IT（Assessing Readability）、T2K（Text-To-Knowledge）等自然语言处理工具，能够实现文本的标注、自动分词、命名实体识别、关系抽取和可视化等多种功能[32]。

3.3 知识挖掘与利用

关联（开放）数据、知识图谱等语义技术和时空分析、关系分析和聚类分析等多维分析可视化方法的发展，使得文化遗产资源的语义知识服务成为可能，不断提升文化遗产知识服务的效果和层次[10]。

3.3.1 关联（开放）数据

在具体的实践过程中，关联数据和开放数据技术应用的最多，关联（开放）数据的实质是以资源描述框架（Resource Description Framework，RDF）数据模型为基础，利用 OWL、SKOS 等工具，将不同内容商提供的非结构化数据或者采用不同标准的结构化数据转换成标准化、结构化的数据[33-34]，实现数据（集）的建模、创建、协调和聚合，具体原理如图 1 所示：

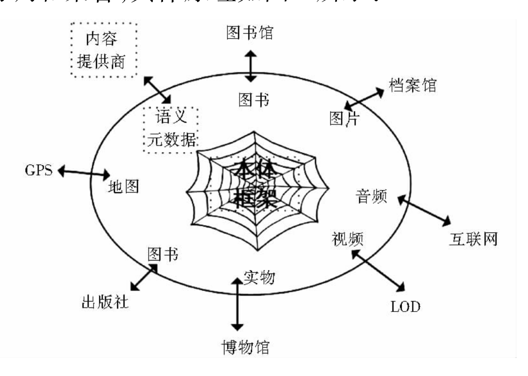


图 1 关联数据原理框架

从实践上来说，关联数据最多实现于图书馆、博物馆、档案馆和美术馆（以下简称"四馆"）等公共文化机构，CultureSampo、BookSampo 和 WarSampo 基于集成语义模型、元数据对齐模型和领域本体模型的 Sampo 关联数据发布模型，分别从文化（如绘画、小说、漫画等资源）、书目（如作者、编辑、出版商等资源）和战争（如战争日记、相册集、回忆录等资源）3 个维度，实现四馆、高校科研机构、出版社等机构的馆藏资源的集成与关联；另外还有欧洲文化遗产门户网站 Europeana、英国博物馆的 ResearchSpace、耶鲁中心的英国艺术项目、美国艺术类文化遗产项目 AAC（American Art Collaborative）均以 CIDOC CRM 框架模型为基础，实现（欧盟）各国四馆、展览馆馆藏资源的关联。

3.3.2 知识图谱

在文化遗产领域，知识图谱常被用来进行概念和语义关系的表示，作为一种数据结构，能够形成一个互联的、分散的全球知识网络[35]，实现大型语料库的联结，为数字信息系统、信息导航、可视化分析、知识发现和语义检索提供支撑。知识图谱的重要意义不仅在于提升查全率、查准率，更能够揭示知识的层级关系，并在此基础上进行多维度的可视化分析。

实践方面，意大利文化遗产项目 ArCo 基于 82 万件文化遗产实体的事件信息、位置信息等信息，构建了 7 个受控词表、1.69 亿个三元组，利用 SPARQL 对 RDF 进行处理并进行数据发布[36]；CultureSampo 构建芬兰历史人物知识图谱，实现人物关系的相互关联；宋代学术语义网络平台构建"宋代学术师承"知识图谱，并开

chinaXiv:202304.00292v1

发出 RelFinder 关系发现工具,将知识图谱转换为能够解释关系属性和自然语言表达的关系实例,基于实例的层次属性实现 <X,Y,Z> 三级关系的精准化语义检索[9]。

3.3.3 多维分析可视化技术

多维分析可视化技术作为知识挖掘的重要手段,主要是利用计算机图形学和图像处理技术,遵照不同的维度,将数字形式的数据转换为图形、图像,予以直观、形象地呈现,目前在文化遗产领域的应用主要包括:聚类分析可视化、时空分析可视化和实体关系分析可视化。

(1)聚类分析可视化。文化遗产领域,常遵照资源的内在相似性将数据集划分为多个类别,根据资源的主题、特征等维度展开聚类分析。当前多数文化遗产项目遵照主题对资源数据进行划分,实现相同主题文化遗产资源的聚合,欧盟文化遗产平台 Europeana 最具代表性,该平台从考古、艺术、时尚、工业遗产、音乐、运动、战争等多个主题角度对欧洲历史进行揭示[4]。在此基础上,部分文化遗产项目以资源特征为基础对文化遗产资源进行组织,并以此为基础进行可视化展示,如 MuseumFinland 基于材料特征、Europeana 基于颜色特征对数据进行可视化分析。

(2)时空分析可视化。时空分析可视化包括时间分析可视化、空间分析可视化和时间－空间分析可视化。

时间分析可视化,即以时间维度为基准进行文化遗产知识的语义组织,实现基于时间轴的时间序列可视化。比如 WarSampo 以时间轴为基础,对战争事件、人物生平等含有时间特征的资源进行可视化展示。

空间分析可视化,即遵照地理空间维度对文化遗产资源进行语义组织,实现基于地图系统的地理空间可视化,当前最受关注的就是地理信息系统(GIS)在数字人文领域的应用。国内目前最具代表性的是浙江大学的地图发布平台,用户可在平台内实现地图的发布、编辑、搜索、查看与共享。文化遗产领域,常见的空间分析可视化的呈现形式包括:分布图(如浙江古塔分布、古代周姓人物分布等)、路线图(如汤显祖行迹图、明代驿站路线图等)。

时间－空间分析可视化,即实现时间分析和空间分析的结合,提供更加完整的时空分析可视化知识服务。芬兰最新的文化遗产语义项目 BiographySampo[37] 能够以地图系统为基础,实现基于时间轴的人物－事件的可视化显示,图 2 为伊利尔·沙里宁的生平事迹的检索结果,显示出了人物在不同时间、不同地点所发生的事件。图 2 中下侧为基于时间轴的事件分布情况,上侧为事件的地理分布情况,二者结合实现事件时间和位置的关联。

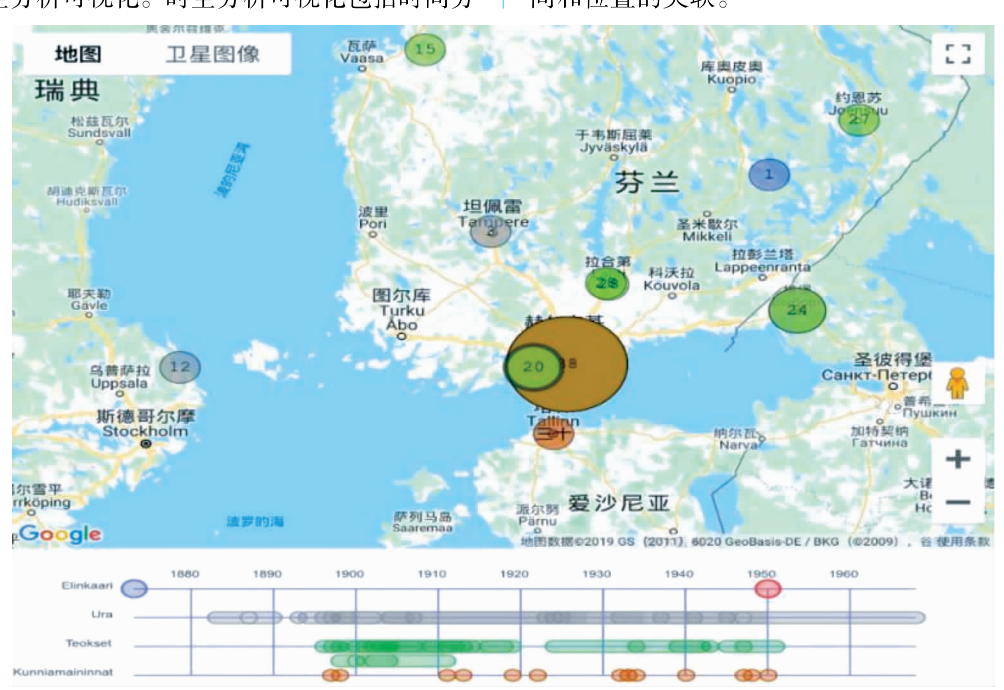


图 2　时间－空间分析可视化示例:BiographySampo[37]

(3)实体关系分析可视化。实体关系分析可视化主要是揭示文化遗产资源中人物的社会关系,其中一个人物或者地点等实体是一个节点,连线表示 2 个实体之间的关系,多个节点、多个连线则构成一个复杂的

chinaXiv:202304.00292v1

ChinaXiv合作期刊

社会网络关系,实现人物与人物/地点等之间关系的语义化揭示。以国内具有代表性的宋代学术语义网络平台为例,平台基于 CBDB 的数据关系整合出学生关系(kv:hasStudent)、子女关系(schema:children)、籍贯(kv:nativePlace)等 22 种实体关系,图 3 为利用宋代学术语义网络平台的关系发现工具 RelFinder 对"王安石–司马光–苏轼"三者之间关系进行揭示[9]。

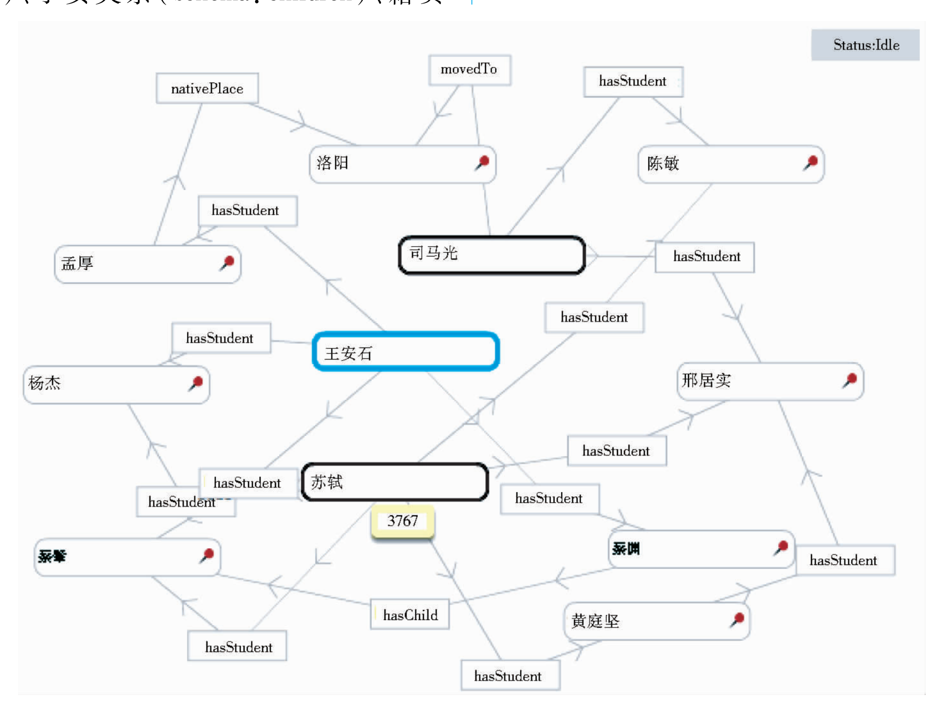


图 3 实体关系分析可视化示例:宋代学术语义网络平台[9]

4 存在的问题及未来发展趋势

4.1 数据互操作问题

由于文化遗产数据具有多学科性和异质性,目前文化遗产组织多是在本体中实现概念和 URI 的简单合并。同时,由于数据来源广泛,使用的数据格式和标准无法统一,很难实现数据的语义互操作。语义网所涉及到的本体、元数据等各项技术,不仅提供了通用标准框架,更为数据互操作提供了通用标准、规范。因此,需要通过语义网技术来优化数据集成和知识再利用,目标应该转向真正的概念集成、本体匹配和语义相关实体的关联,还应该考虑到数据是否是由不同的方法或具有不同科学背景的用户创建的。同时,利用多学科信息进行模糊推理的语义技术规则和概率描述逻辑,以及对动态演化数据的推理是必要的,从而实现深层语义挖掘和多维信息的概念融合与集成。

4.2 本体标准化问题

本体构建一直是文化遗产领域的关注热点,目前多数项目以 CIDOC CRM 为基础本体模型进行扩展或改造,能够提升语义组织效率,但是也会导致语法缺陷,概念和模型缺陷和不一致性等语义问题。另外,不同机构或领域构建的本体使用不同的主题词表或标准,也会导致相互孤立、互不兼容等语义问题。为解决此问题,必须要从构建主体和开放获取两个维度展开,即:扩大本体构建的主体范围,让更多的个人、组织机构加入到本体的建设过程当中;增强本体框架的开发获取水平,实现本体共享的同时,也能够加强本体标准化建设。此外,本体的映射和对齐、基于基本逻辑原则的本体共享也能够实现本体标准化,也有助于增强不同领域本体间的语义互操作。

4.3 多语言知识表达问题

本文在研究过程中发现文化遗产呈现明显的多语言性质,且大多数机构用本地语言存储数据。开放数据运动鼓励文化遗产机构向公众提供其数据,即使这一做法被广泛接受,也很难确保机构用本地语言以外的其他语言发布其数据集合。为了实现文化遗产数据在世界范围内的共享、使用,就迫切需要在系统、模型开发的源头就开始注重多语言的问题,比如本体多语言化、元数据标准多语言化和多语言的知识表示、获取和翻译。因此,在未来文化遗产项目的设计开发过程中,本地化和语言技术处理、翻译与表示和自然多语言数据管理等技术将会进一步发展。

chinaXiv:202304.00292v1

4.4 个性化语义问题

个性化语义是表达主观意见和主观推理的必要条件。具有特定属性的文化遗产对象,对于背景、兴趣和目的不同的人来说,往往具有不同的审美意义或解释意义。同样,有争议的文化遗产也可能会受到褒贬不一的评价和感受。解决此问题,关键在于为用户提供个性化语义服务,包括个性化主页、语义超链接、个性化语义检索、个性化语义导航等[38]。因此,语义推理技术将会成为未来的发展趋势之一,即根据文化遗产对象之间的语义关系进行推理,无论是使用描述逻辑,还是语义知识提取,都能够派生出不同的语义并相应地进行个性化知识服务。

4.5 自动化工具构建问题

目前,国内外仅有少数项目能够在对文化遗产进行语义组织的过程中进行自动化工具开发。在文化遗产语义组织过程中,文化遗产机构开始应该使用语义技术处理数据,并将其作为关联数据发布,同时应该开发一些使用简便,且集成技术流程、能够处理大量数据的自动化工具,比如自动标注、自动分词、可视化分析工具等。尽管已经存在一些相应的工具,但是其自动化程度较低,如果要增强文化遗产组织对语义网的利用程度,还需要进一步提升工具的自动化和便利程度。

4.6 数据版权问题

数字人文和开放数据运动的盛行,目前数字人文资源的规模日益庞大。数据版权问题没有得到完全解决,将成为博物馆、图书馆等机构进行信息分布和共享的阻碍。为解决此问题,可以借鉴现有的知识产权保护相应的措施,从法律、制度、执行许可证等多个层面保护数据版权。同时,引入元数据技术与标准、区块链技术,亦能对数据版权起到保护作用。

5 总结

本文以国内外典型文化遗产项目的调研数据为基础,从语义组织方式和知识服务与工具两个方面对文化遗产项目语义组织的研究现状进行梳理。同时,本文基于知识建模、知识抽取和知识挖掘与利用 3 个维度,对文化遗产语义组织关键技术进行剖析,具体包括本体、元数据、关联(开放)数据、知识图谱等技术。通过上述分析,本文发现文化遗产项目的语义组织在未来的发展过程中仍面临诸多挑战:比如增强数据的深度互操作,实现领域本体标准化、系统性和可重用性,满足用户对于知识服务的个性化、多元化和多层次需求,加大自动化、易用工具的开发力度,增强文化遗产数据的版权保护等问题。

参考文献:

[1] FULLER M. An account of the structure of the China biographical database[EB/OL]. [2019－11－10]. https://projects. iq. harvard. edu/chinesecbdb/% E8% B3% 87% E6% 96% 99% E5% BA% AB% E6%9E% B6% E6% A7%8B.
[2] WarSampo: Finnish world war ii on the semantic Web [EB/OL]. [2019－12－03]. https://seco. cs. aalto. fi/projects/sotasampo/en/.
[3] HYVÖNEN E, JUNNILA M, KETTULA S, et al. Finnish museums on the semantic Web: the user's perspective on museumfinland[C]// Proceedings of Museums and the Web 2004. Arlington,2004.
[4] About EUROPEANA[EB/OL]. [2020－01－30]. https://www. europeana. eu/portal/en/about. html.
[5] CBDB 资料来源与涵盖范围[EB/OL]. [2020－01－30]. https://projects. iq. harvard. edu/chinesecbdb/% E8% B3% 87% E6%96% 99% E4% BE% 86% E6% BA% 90% E8% 88% 87% E6% B6% B5% E8%93%8B% E7% AF%84% E5%9C%8D.
[6] 龚延明. 重构宋代四万进士档案——浙大宋学中心龚延明、祖慧《宋代登科总录》(14 册)介绍[EB/OL]. [2020－01－30]. http://rwsk. zju. edu. cn/2015/1112/c2039a170378/page. htm.
[7] 傅君勱, FULLER M A. 中国历代人物传记资料库用户指南(中文版)[EB/OL]. [2019－10－25]. http://172. 16. 20. 58/cache/2/03/projects. iq. harvard. edu/6ed82dd42b570535d90551ae3c305d66/cbdb_users_guide_ch_170126. pdf.
[8] 陈佩诗. 资讯技术与历史文献分析[D]. 台北:台湾大学,2011.
[9] KVISION. 宋代学术语义网络——关系发现(RELFINDER)[EB/OL]. [2019－10－25]. http://dh. kvlab. org/cbdb_kg/.
[10] 左丹,欧石燕. 人文信息资源语义描述、语义组织研究与实践述评[J]. 图书馆论坛,2019,39(8):21－31.
[11]《历代进士登科数据库》产品介绍[EB/OL]. [2019－10－15]. http://examination. ancientbooks. cn/docDengke/dkHelp. jspx.
[12] Getty vocabularies[EB/OL]. [2019－10－20]. http://www. getty. edu/research/tools/vocabularies/index. html.
[13] YSA——General Finnish thesaurus[EB/OL]. [2019－10－20]. https://www. kansalliskirjasto. fi/en/node/167.
[14] 敦煌壁画主题词表项目介绍[EB/OL]. [2019－10－20]. http://dh. whu. edu. cn/dhvocab/dhresource/html/intro. html.
[15] 刘宗田. 事件本体体系结构[J]. 计算机科学,2012,39(2):45.
[16] 王楠. 历史人物本体构建及其查询推理研究[D]. 桂林:广西师范大学,2017.
[17] CIDOC CRM[EB/OL]. [2019－10－20]. http://www. cidoc-crm. org/.
[18] General EDM factsheet and presentation[EB/OL]. [2019－10－20]. https://pro. europeana. eu/resources/standardization-tools/edm-documentation.

chinaXiv:202304.00292v1

[19] Bibliographic ontology[EB/OL]. [2019-10-20]. https://en.m.wikipedia.org/wiki/Bibliographic_Ontology.

[20] STIVERS C. The ontology of HOPE in dark times[J]. Administrative theory & praxis, 2008, 30(2): 225-239.

[21] VAN H W R, MALAISÉ V, SEGERS R, et al. Design and use of the simple event model[J]. Social science electronic publishing, 2011, 9(2): 128-136.

[22] What is arches[EB/OL]. [2019-10-20]. https://www.archesproject.org/what-is-arches/.

[23] ResearchSpace[EB/OL]. [2019-10-20]. https://www.researchspace.org.

[24] About CLAROS[EB/OL]. [2019-10-20]. http://explore.clarosnet.org/XDB/ASP/clarosHome/about.html.

[25] About MIMO[EB/OL]. [2019-10-20]. http://www.mimo-international.com/MIMO/about-mimo.aspx.

[26] Carare association[EB/OL]. [2019-10-20]. https://www.carare.eu/about/.

[27] HYVÖNEN E. Cultural heritage linked data on the semantic web: three case studies using the sampo model[C]// VIII Encounter of documentation centres of contemporary art: open linked data and integral management of information in cultural centres. Vitoria-Gasteiz: Artium, 2016: 19-20.

[28] RUOTSALO T, HYVÖNEN E. An event-based approach for semantic metadata interoperability[C]// The international semantic web conference. Heidelberg: Springer, 2007: 409-422.

[29] 王红霞,苏新宁. 基于元数据的电子政务信息资源组织模式[J]. 情报理论与实践,2007,30(1):116-121.

[30] 王萍,黄新平. 基于关联数据的数字文化资源语义融合方法研究[J]. 图书情报工作,2016,60(12):29-37.

[31] HYVÖNEN E, LINDQUIST T, TÖRNROORS J, et al. History on the semantic web as linked data——an event gazetteer and timeline for the world war i[EB/OL]. [2019-10-25]. https://seco.cs.aalto.fi/publications/2012/hyvonen-et-al-ww1-cidoc-2012.pdf.

[32] T2K(Text-To-Knowledge)[EB/OL]. [2019-10-25]. http://www.italianlp.it/demo/t2k-text-to-knowledge/.

[33] 夏翠娟,刘炜,赵亮,等. 关联数据发布技术及其实现——以Drupal为例[J]. 中国图书馆学报,2012(1):51-59.

[34] BIZER C, HEATH T, BERNERS L T. Linked data: the story so far[C]// Semantic services, interoperability and Web applications: emerging concepts. Hershey: IGI Global, 2011: 205-227.

[35] HASLHOFER B, ISAAC A, SIMON R. Knowledge graphs in the libraries and digital humanities domain[EB/OL]. [2019-10-25]. https://arxiv.org/pdf/1803.03198.pdf.

[36] GROUNDAI. ArCo: the Italian cultural heritage knowledge graph[EB/OL]. [2019-10-25]. https://www.groundai.com/project/arco-the-italian-cultural-heritage-knowledge-graph/1.

[37] BiographySampo[EB/OL]. [2019-10-25]. http://biografiasampo.fi/henkilo/p992/kartat.

[38] 鞠彦辉. 个性化语义门户知识组织研究[J]. 情报科学,2014(3):53-56,67.

作者贡献说明:

李章超:论文撰写与修改;

何琳:论文选题与框架设计,提出论文修改建议。

Case Study on Semantic Organization of Cultural Heritage

Li Zhangchao He Lin

College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] This paper focuses on the development of semantic organization of cultural heritage has important reference value for the study of cultural heritage in China. [**Method/process**] This paper adopted the method of systematic research, case analysis and statistical analysis, based on the survey data summaries, combed the research status of semantic organization of cultural heritage projects from semantic organizations and knowledge services and tools, and analyzed the key technologies of semantic organization of cultural heritage from 3 aspects: knowledge modeling, knowledge extraction, knowledge mining and knowledge utilization. [**Result/conclusion**] The research finds the keys to the development of semantic organization of cultural heritage are data interoperability, domain ontology standardization, personalized semantics, automation tools, data copyright.

Keywords: cultural heritage semantic organization ontology knowledge services

chinaXiv:202304.00292v1